ACS Combinatorial Science

Research Article

pubs.acs.org/acscombsci

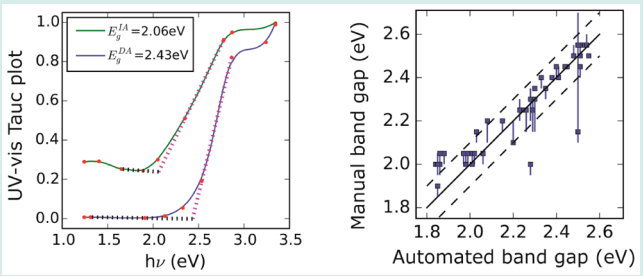
High Throughput Light Absorber Discovery, Part 1: An Algorithm for Automated Tauc Analysis

Santosh K. Suram, Paul F. Newhouse, and John M. Gregoire*

Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena California 91125, United States

Supporting Information

ABSTRACT: High-throughput experimentation provides efficient mapping of composition−property relationships, and its implementation for the discovery of optical materials enables advancements in solar energy and other technologies. In a high throughput pipeline, automated data processing algorithms are often required to match experimental throughput, and we present an automated Tauc analysis algorithm for estimating band gap energies from optical spectroscopy data. The algorithm mimics the judgment of an expert scientist, which is demonstrated through its application to a variety of high throughput spectroscopy data, including the identification of indirect or direct band gaps in Fe_2O_3, $Cu_2V_2O_7$, and $BiVO_4$. The applicability of the algorithm to estimate a range of band gap energies for various materials is demonstrated by a comparison of direct-allowed band gaps estimated by expert scientists and by automated algorithm for 60 optical spectra.



KEYWORDS: *high-throughput screening, combinatorial science, band gap, UV−vis spectroscopy, optical spectroscopy, solar fuels*

INTRODUCTION

Modern technologies, such as photovoltaics, solar fuels, electrochromic devices, window coatings, and pigments, require materials with specific optical properties. High-throughput (HiTp) and combinatorial materials science approaches have been successfully applied for the discovery of high-performance optical materials.[1] Efficient synthesis techniques employing HiTp approaches provide access to high-order, for example, ternary and quaternary, composition spaces, which have been sparsely explored for the discovery of functional materials. Applying HiTp optical characterization on such material libraries produces vast quantities of spectral data; and some optical parameters, such as average reflectivity over the visible range, can be readily calculated.[2] Other performance metrics are model-dependent and are traditionally extracted through manual analysis. In this article, we focus on the data analytic aspects of estimating band gap energies from HiTp transmission and reflection responses of combinatorial material libraries under ultraviolet−visible (UV−vis) illumination. We demonstrate our approach in the context of HiTp screening of composition libraries for the discovery of light absorbers for solar energy applications.

While the optimal band gap of a light absorber for photovoltaic or solar fuels applications varies with device design, the identification of new semiconductors with band gap energies in the visible to near-ultraviolet range is important for advancing these technologies.[3] The band gap energy is the primary performance metric for qualifying a material as a candidate for solar energy applications, motivating rapid measurement of band gap energy for accelerating the discovery of solar absorber materials. In particular, the HiTp methods must readily identify the band gap energy of light absorber phases in a composition library and any systematic variation in band gap energy with respect to composition, a common manifestation of composition alloying within a light absorber phase.[4]

High-throughput optical property measurements typically rely on estimation of properties from transmission and reflectance responses of materials under illumination. We recently described an on-the-fly high-throughput spectroscopy instrument capable of measuring the transmission and total reflection responses (TR instrument) at a throughput better than 1 sample per second[5] and have also adapted this technique for diffuse reflectance measurements (DR instrument). Estimation of band gap energy from these high throughput measurements is an arduous task when using traditional manual analysis by an expert scientist, limiting the efficacy of high-throughput data acquisition. In this article, we present a constrained piecewise linear fitting based algorithm that allows rapid estimation of band gap energy (typically, <0.5 s per spectrum on an Intel Core i7-3770 CPU @ 3.4 GHz, 16 GB RAM, 64-bit Windows 7 OS). We build intuitive parameters into the algorithm to mimic an expert scientist's judgment so that absorption spectra that do not conclusively identify a band gap are flagged, and the band gap energy is calculated for the remaining samples. We demonstrate this approach by applying the algorithm to the characterization of various metal oxides using both the TR and DR measurement techniques.

Received: April 9, 2016
Revised: September 8, 2016
Published: September 23, 2016

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673−681

Tauc Analysis and Existing Algorithms. In both automated and manual estimation of the band gap energy from an optical absorption spectrum, researchers typically employ the Tauc relationship:[6]

$$\alpha h\nu \propto (h\nu - E_g)^n \quad (1)$$

where h is the Planck's constant, ν is the frequency of light, and E_g is the band gap energy. The value of the exponent n is related to the electronic nature of the band gap, with values of 3, 2, 3/2, and 1/2 corresponding to indirect forbidden (IF), indirect allowed (IA), direct forbidden (DF), and direct allowed (DA) transitions, respectively. This approximate relationship between photon energy and absorption is particularly relevant when the absorption coefficient (α) exceeds 10^4 cm^{-1} for photon energies approximately 1 eV above the band gap energy.[6] This condition essentially requires the semiconductor under investigation to be a strong absorber, and for the purpose of identifying materials for solar absorption applications, materials that do not meet the strong absorber criterion are not of primary interest.

The band gap energy, E_g, is typically determined through inspection of $(\alpha h\nu)^{1/n}$ vs $h\nu$ plots (also called Tauc plots),[7] where the linear trend given by eq 1 is modeled as the tangent of the Tauc plot near the point of maximum slope if the Tauc plot contains a sufficiently linear region.[8] This linear tangent is then extrapolated to the point where $(\alpha h\nu)^{1/n}$ is 0 or a small value provided by modeling the baseline signal.

It is important to note that although the Tauc approach was derived using the parabolic band representation of localized energy states for amorphous materials, it typically remains valid for polycrystalline materials using parabolic band approximation based on Maxwell–Boltzmann statistics.[9] However, in case of defective or degenerate semiconductors, the band gap observed in optical absorption measurements (apparent band gap) could differ from the fundamental band gap of the semiconductor due to effects such as Burstein–Moss shifts,[10] band gap renormalization due to electron–electron and electron–ion interactions,[11] and formation of dopant levels within the band gap. While empirical estimation of the fundamental band gap from optical spectra of degenerate semiconductors is an active area of research,[9] the intentional modification of optical absorption by tailoring defect chemistry is also a promising strategy,[12] and for the purposes of semiconductor discovery for solar absorption, the apparent band gap is a suitable property to map in high throughput studies. Given the high throughput implementation of Tauc analysis in the present work, each reported band gap is understood to be the apparent band gap of the material in its thin film (and likely defective) form.

Since manual inspection of Tauc plots to identify the point of maximum slope is prohibitive for rapid estimation of band gaps, several semiautomated algorithms for Tauc analysis of absorption spectra have been reported. Anderson et al.[13] provide composition-band gap mapping by providing user-specified ranges of the photon-energy and Tauc-value ($(\alpha h\nu)^{1/n}$) to define a portion of the Tauc plot for linear regression. The extrapolation of the fitted Tauc line to the $(\alpha h\nu)^{1/n} = 0$ line provides the estimate of E_g. Ghobadi et al.[14] use a similar approach to determine band gap as a function of particle size in CdSe nanostructured thin films. Such algorithms are not amenable to automated analysis on material libraries in which the band gap energy may span a wide energy range. To overcome the need for user-specified partitions of the Tauc plot, the maxima in the first-derivative of $(\alpha h\nu)^{1/n}$ vs $h\nu$ may be used to identify data points whose tangents represent the Tauc line. However, this approach is sensitive to noise and local optical features, as demonstrated and discussed by Escobedo Morales et al.[8]

It is also important to recognize that absorption spectra may contain small contributions from nonlinear effects (e.g., plasmonic scattering) and subgap absorption tails due to defect states or intraband absorption.[15] For IA, IF, and DF gaps, the large value of n in eq 1 amplifies the importance of these contributions to the absorption signal, possibly resulting in a skewed band gap energy upon extrapolation of the fitted Tauc line to $(\alpha h\nu)^{1/n} = 0$. To address these issues, we developed an analysis algorithm that partitions the Tauc plot into a series of linear segments and then identifies the linear segment most representative of the Tauc relationship based on the energy ranges, Tauc value ranges, and Tauc slopes of the line segments. The algorithm also identifies a line segment to model the subgap absorption baseline, and the intersection of the two lines provides the estimation of E_g. We have encoded this process into an automated algorithm, and below we demonstrate its ability to interpret Tauc spectra from (a) 49 duplicate α-Fe_2O_3 samples using TR data, (b) α-$Cu_2V_2O_7$ using TR data, (c) monoclinic-$BiVO_4$ using DR data, and (d) a biphasic sample using DR data.

Some limitations of high throughput band gap mapping and opportunities for enhancing high throughput materials discovery are also discussed. The algorithm contains several free parameters whose values have been chosen through manual consideration of thousands of Tauc spectra. The ability of the algorithm to mimic expert judgment is demonstrated by comparing its analysis of various oxide and sulfide thin films to that of expert scientists. This automated extraction of band gap energy is an enabling technology for light absorber discovery and more generally for combinatorial materials science, as demonstrated in part 2 of this manuscript series.

■ ALGORITHM AND DISCUSSION

Generation of Tauc Property Spectra. In this section, we describe processing of optical data to generate the Tauc spectra (eq 1) on which the band gap estimation algorithm is applied. For the TR instrument, the measured transmission and reflection spectra are scaled by a reference transmission spectrum obtained on a bare substrate and a Spectralon white cube, respectively. The resulting fractional transmission (T) and fractional reflection (R) spectra are processed for outlier removal and noise filtering using a fourth-degree polynomial Savitzky–Golay filter and photon energy filtering window of 45 nm.[16] The absorption coefficient of the sample and the transmission are related according to the Beer–Lambert equation ($\alpha L = -\ln T$).[17] With an approximation that the fractional intensity of light that has the opportunity to be absorbed by the sample is given by $1 - R$, the Beer–Lambert equation is modified[18] as follows:

$$\alpha L = -\ln\left(\frac{T}{1-R}\right) \quad (2)$$

where α is the absorption coefficient and L is the optical path length through the sample thickness.

For the DR instrument, a Spectralon white reference is used to calculate the fractional diffuse reflection (DR) spectra, which is then smoothed using Savitzky–Golay filtering. The

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673–681

absorption coefficient (α) can be expressed in terms of DR according to the Kubelka–Munk radiative transfer model:[19]

$$f(R) = \frac{1 - DR^2}{2DR} = \frac{\alpha}{s} \tag{3}$$

where s is the spectral scattering factor. We adopt the common assumption that s is a constant.[20] This approximation is a practical necessity for HiTp analysis and introduces the possibility of systematic errors in the extracted band gap from DR data, which can be addressed for representative samples using detailed optical characterization techniques such as spectroscopic ellipsometry.[21]

With this data processing, both TR and DR instruments yield optical spectra that are proportional to α, leading to an instrument-specific definition of the quantity β, which is defined as αL (eq 2) for the TR instrument and α/s (eq 3) for DR instrument. To apply eq 1, we define the family of Tauc property spectra as

$$TP = (\beta h\nu)^{1/n}/\max((\beta h\nu)^{1/n}) \tag{4}$$

where n is the Tauc exponent used in eq 1, which in the present work is limited to IA ($n = 2$) and DA ($n = 1/2$) band gaps. The corresponding Tauc property spectra are referred to as TP^{IA} and TP^{DA}, respectively, which are interpreted to ascertain the respective band gap energies, E_g^{IA} and E_g^{DA}. Each Tauc spectrum is scaled by its maximum value (see eq 4), making the proportionality constants in eqs 2 and 3 inconsequential and enabling the parameters for the band gap estimation algorithm to be generalized for all types of TP. We note that this scaling makes the TP spectrum unitless, and since TP is analyzed as a function of $h\nu$, slopes of this spectrum have units of eV^{-1}.

Constrained Piecewise Linear Fitting. TP spectra are fit with $k + 1$ linear segments using k nonterminal nodes, and the value of k is chosen so that the piecewise linear regression model routinely fits the data with a coefficient of determination in excess of 0.99, and we choose $k = 6$ upon examination of several TP spectra. The linear fit is optimized by minimizing the loss function, which is defined as normalized sum of squared errors, that is, $\sum((TP - TP_{fit})/TP)^2$. The division by TP in this objective function promotes a better fit for the baseline absorption and mitigates sensitivity to noise in the higher absorption region of TP. In particular, this normalization has an important consequence for fitting the nonlinear onset of absorption near the band gap energy. With a typical least-squares objective function, this nonlinear region can be fit by a single line with little penalty since TP is small at the onset of absorption. Our modification to the traditional objective function amplifies the importance of this absorption onset region such that it is usually modeled with multiple linear segments, which promotes the identification of baseline signals that mimic those identified by expert scientists. During the segmented linear fitting procedure, the energy offset between successive nodes is constrained to be greater than $p_1 = 0.05$ eV, which discourages overfitting of local features (typically noise). Additionally, noise in TP at lower and higher energy limits of the instrument can cause the first and last linear segments to misrepresent the data. Therefore, these segments are retained for further analysis only if their energy range is greater than a certain minimum, which we chose to be twice the minimum offset between nodes ($p_2 = 2p_1 = 0.1$ eV).

As discussed below, the band gap estimation algorithm can be sensitive to an excess of line segments in the fitting model, and while $k = 6$ is often necessary, some TP spectra require fewer line segments. While an iterative fitting routine with different values of k could be used to identify the requisite number of line segments, we have developed an alternate strategy that focuses on removing the primary artifact of overfitting, that is, dissecting the linear Tauc region into multiple line segments. If a region of TP is nearly linear and would be considered by an expert analyst to be the linear Tauc region, that entire region should be modeled as a single line segment in the fitting result. This performance can be achieved by starting with an overfit line segment model followed by merging of neighboring line segments that are sufficiently similar. We implement this strategy by merging neighboring line segments that meet the following criterion:

$$\delta_{i,\text{merge}} < p_3(\Delta(TP)_i + \Delta(TP)_{i+1}), \quad i \in [1, k] \tag{5}$$

where $\Delta(TP)_i$ and $\Delta(TP)_{i+1}$ are the extent of TP values traversed by the ith and $(i + 1)$th linear segments, respectively. The quantity $\delta_{i,\text{merge}}$ is the maximum difference in TP between this pair of neighboring line segments and the line segment created by merging these line segments into a single line segment that connects their extreme end points. This quantity is illustrated in Figure 1, which demonstrates that a merged line segment sufficiently represents a pair of neighboring line segments if $\delta_{i,\text{merge}}$ is small. This merging step is repeated iteratively until no neighboring line segments meet the above criterion, and we find that the resulting line segment model of TP closely mimics the judgment of experts, particularly with the parameter $p_3 = 0.1$. Since the number of line segments may be lower than k after this merging procedure, we define the postmerge number of line segments as ns.

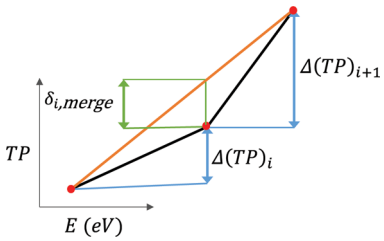


Figure 1. Geometric illustration of the parameters considered for merging line segments in eq 5. Neighboring line segments (black lines) are merged as a single line (orange line) if $\delta_{i,\text{merge}}$ is sufficiently small.

Thin-Film Interference Effects. Thin film interference is a well-known phenomenon in which light reflected at the film–air and film–substrate interfaces of a thin film interfere to generate oscillations in the reflection and transmission spectra. In the case of optical spectra obtained from the TR instrument, the term $T/(1 - R)$ in eq 2 provides interference-free determination of optical absorption coefficient.[22] For DR spectra, the application of Kubelka–Munk radiative transfer model (eq 3) assumes that a vanishingly small fraction of incident light is diffusely reflected from the film–substrate interface; and consequently the model is not applicable to DR spectra exhibiting interference oscillations. For the present purpose of solar light absorber discovery, we identify and disregard such spectra because thin-film interference oscillations are generally observed in weakly absorbing samples. Prior to application of our band gap estimation algorithm, we filter out samples whose optical spectra have significant

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673–681

contributions from interference patterns using a minimum allowed slope (MAS) criterion (eq 6):

$$\min(S_i) > p_4 \quad \forall\, i \in [1, ns] \tag{6}$$

where S_i is the slope of the ith line segment. *TP* spectra that fail this criterion are excluded from further analysis, and we observe that $p_4 = -2$ eV^{-1} provides robust identification of spectra with significant thin-film interference effects. While alternate methods for identifying interference fringes can be employed, the MAS criterion for the piecewise fit linear segments provides interference detection with less sensitivity to noise than, for example, peak identification methods.

Band Gap Estimation. With a line segment model of *TP* in hand, the algorithm proceeds by identifying both the line segment that best represents the above-gap linear Tauc region (the "Tauc line segment", see eq 1) and the line segment that best represents the below-gap baseline. Given a Tauc line segment and baseline segment, the abscissa of the intersection of these lines represents the corresponding band gap energy. It is worth noting that multiple pairs of absorption onset and baseline linear segments could be found in a single *TP* spectrum, for example, in the case of a multiphase system. The algorithm is thus generalized to permit the identification of an arbitrary number of band gaps, although this number is practically limited to 2 or 3 when the algorithm is seeded with $k = 6$ nonterminal nodes.

We proceed by defining and describing the set of inequalities for evaluating whether a given line segment (index τ) from the fitting model is a Tauc line segment with a well-defined baseline. A primary criterion set ($\mathbf{C}_1$) is used to evaluate each nonterminal line segment ($\tau \in [2, ns-1]$) by comparing its properties to those of the preceding (lower photon energy, index $\tau - 1$) and succeeding (higher photon energy, index $\tau + 1$) line segments. A second, modified criterion set ($\mathbf{C}_2$) is used to evaluate the last line segment ($\tau = ns$).

Tauc Line Segment Identification. Criterion set 1 ($\mathbf{C}_1$): To mimic an expert's estimation of the Tauc line as the tangent of *TP* at the point of maximum slope, the slope of Tauc line segment τ must exceed that of its preceding and succeeding line segments. To elucidate the prudence of these criteria, we consider the relationship of the Tauc line segment with the lower-energy absorption edge, which is typically manifested as an exponential tail.[6] As a result, *TP* has positive curvature for photon energies just below the linear Tauc region. Attributing a region within this absorption tail to a Tauc line segment can lead to significant underestimation of the band gap. By requiring that the slope of the Tauc line segment is greater than those of the preceding and succeeding line segments, the identified Tauc line will not be in the absorption tail, providing the best estimate of the band gap energy. Since the line segment preceding the Tauc line segment is part of the absorption tail, it should have a positive or approximately zero slope, leading to an additional criterion that the slope of the preceding line segment exceed a minimum value of $p_5 = -0.05$ eV^{-1}. In addition, line segment τ should capture at-least a minimum *TP* difference of $p_6 = 0.1$ to ensure that the Tauc line segment region represents a substantial portion of the normalized *TP*. Compared to the alternate strategy of identifying the maximum slope point in *TP* and constructing the Tauc line from the corresponding tangent, we find that combining the line segment merging strategy with the following set of inequalities provides an algorithm that is less susceptible to noise and more consistent with an expert's analysis:

$$p_5 < S_{\tau-1} < S_\tau \tag{7a}$$

$$S_\tau > S_{\tau+1} \tag{7b}$$

$$S_\tau > 0 \tag{7c}$$

$$\Delta(TP)_\tau > p_6 \tag{7d}$$

Criterion Set 2 ($\mathbf{C}_2$): Evaluating the terminal line segment as a Tauc line segment is necessary for detecting a band gap near the upper limit of the photon energy measurement range. Since a subsequent line segment is not available for evaluation of the inequalities in eq 7, the applicable subset of criteria are adopted and the $\Delta(TP)$ criterion is made more restrictive. An additional criterion that the photon energy range (ΔE_{ns}) spanned by the line segment is sufficiently large is added to eliminate artifacts from sharp transients in the absorption signal that can arise due to poor signal-to-noise at the highest photon energies, where the lamp intensity is weak. The resulting set of inequalities for $\tau = ns$ are summarized as

$$p_5 < S_{ns-1} < S_{ns} \tag{8a}$$

$$S_{ns} > 0 \tag{8b}$$

$$\Delta(TP)_{ns} > 2p_6 \tag{8c}$$

$$\Delta E_{ns} > 2p_2 \tag{8d}$$

Baseline Segment Identification. For a Tauc line segment (line segment index τ), the corresponding baseline segment (line segment index B) is defined as the lowest energy linear segment that precedes the Tauc line segment and has a slope that is sufficiently less than the Tauc line segment but above the minimum value defined in $\mathbf{C}_1$. Further, a constraint that slope of the linear segments monotonically increases from baseline segment up to the Tauc line segment is added, as is expected for the absorption onset.

$$E_B < E_\tau \tag{9a}$$

$$(S_\tau - S_B) > p_7 \tag{9b}$$

$$S_B > p_5 \tag{9c}$$

$$S_{j+1} > S_j \quad \forall\, j \in [B, \tau - 1] \tag{9d}$$

where E_i is the high energy terminal value for line segment i, and $p_7 = 0.2$ eV^{-1}.

This criteria set promotes identification of a baseline that precedes the nonlinear absorption tail described above, for using this tail as a baseline would result in overestimation of the band gap energy. If a baseline segment cannot be identified for a given Tauc line segment, the Tauc line segment is not used for band gap estimation. When multiple Tauc line segments are identified in a single *TP*, the baseline line segment for a given Tauc line segment must succeed any lower-energy Tauc line segments.

These criteria complete the band gap estimation algorithm;, and a summary of the analysis flow and parameters involved in each process are provided in Figure 2 and Table 1, respectively. In the following section, we discuss the application of this algorithm to a wide range of light absorbers.

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673−681

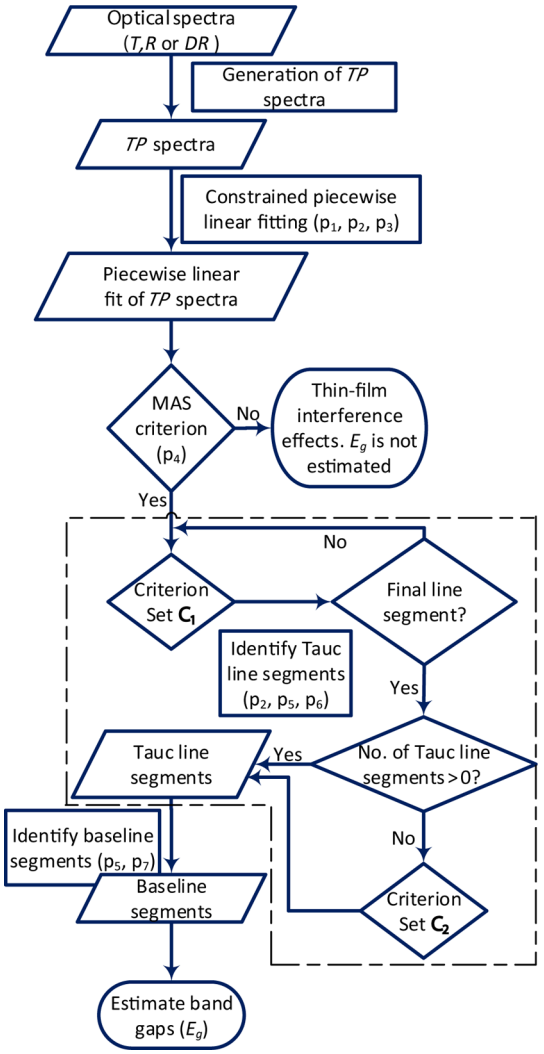


Figure 2. Flowchart of the automated band gap estimation algorithm. The steps employing user-defined parameters are noted, and the dashed line encloses the subroutine for identifying the Tauc line segment(s).

RESULTS AND DISCUSSIONS

To demonstrate the algorithm and verify the estimation of band gap energy, we apply the algorithm to the characterization of phase-pure metal oxides. As a first example, we apply the algorithm described above to optical spectra obtained using TR instrument on 49 duplicate 1 mm^2 Fe_2O_3 thin film samples synthesized on a FTO coated glass substrate via inkjet printing.[5] Figure 3 depicts the initial data processing for a representative sample where the fraction T and R spectra (Figure 3a) were used to calculate β, the absorption coefficient up to a factor of the film thickness (eq 2, Figure 3b). To investigate the presence of both a direct (DA) and an indirect (IA) band gap, the respective TP spectra were calculated using eq 4 (Figure 3c).

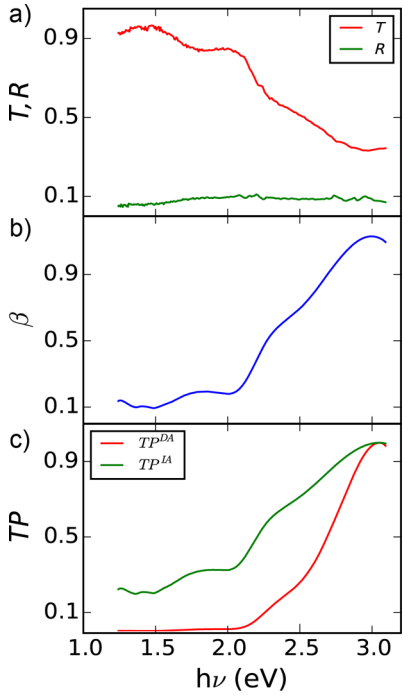


Figure 3. (a) Fractional transmission (T) and reflection (R) spectra for a representative Fe_2O_3 sample; (b) spectrum proportional to absorption coefficient, as derived from T,R using eq 2; (c) direct-allowed (DA) and indirect-allowed (IA) TP spectra. Each TP spectrum is analyzed using the constrained piecewise linear fitting algorithm to identify the band gap energy.

The TP spectra for IA and DA were analyzed using the automated algorithm, resulting in positive detection of a single

Table 1. Details of the Parameters Involved in the Automated Band-Gap Estimation Algorithm

parameter	description	value	reasoning
p_1	minimum offset between successive nodes of piecewise linear fits.	0.05 eV	discourage overfitting of local features by separating linear segment nodes along the energy axis.
p_2	minimum energy span for first and last linear segments.	0.1 eV	similar to p_1 with increased value to avoid overfitting of transients at extremes of the spectra.
p_3	fraction of TP value traversed by neighboring line segments to determine if they should be merged.	0.1	mimic the level of linear smoothing that is visually performed by an expert analyst, which scales with the slope of the TP in the surrounding region.
p_4	determines the minimum allowed slope of the line segments.	-2 eV^{-1}	provides robust identification of spectra with significant thin-film interference effects.
p_5	minimum slope for a line segment preceding the Tauc line segment up to and including the corresponding baseline segment.	-0.05 eV^{-1}	encodes the physical requirement that a Tauc line segment is preceded by line segments that are either part of an absorption tail or a baseline and should have a shallow or positive slope.
p_6	minimum Δ(TP) captured by a Tauc line segment	0.1	encourages the identification of linear features that describe a significant fraction of the TP signal.
p_7	minimum difference in slope between a Tauc line segment and its corresponding baseline segment	0.2 eV^{-1}	requires band gap transition to yield a substantial increase in TP slope and discourages false band gap identification from subgap absorption, for example, from defect states or intraband transitions.

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673−681

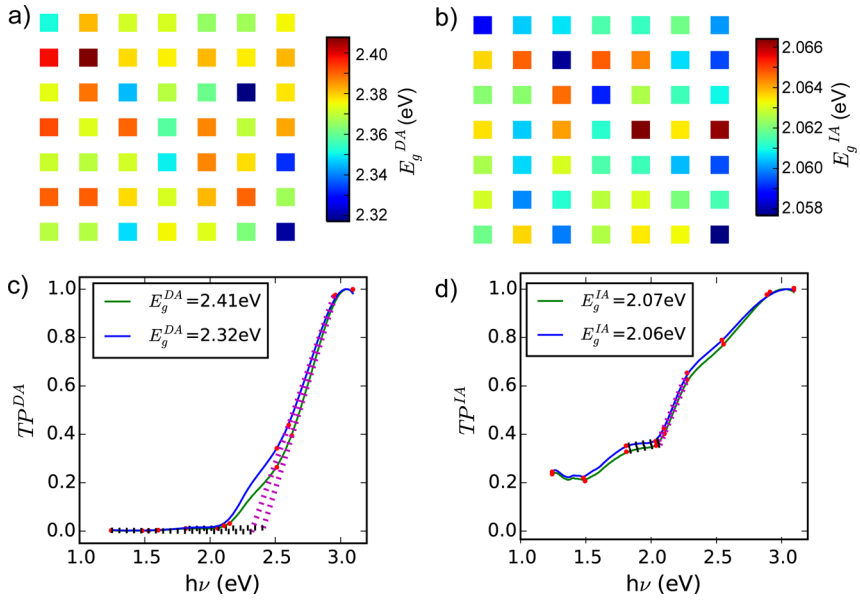


Figure 4. False color mapping of (a) E_g^{DA} and (b) E_g^{IA} for 49 duplicate Fe_2O_3 samples. The 1 mm^2 samples were organized as a 7 × 7 square grid with a 2 mm pitch, and the calculated band gap energies are plotted according to this physical layout. The *TP* spectra for samples exhibiting the smallest and largest estimates of (c) E_g^{DA} and (d) E_g^{IA} are also shown. For each plot, the end points of the piecewise linear fit are highlighted as red circles. The Tauc line segment (dotted magenta), baseline segment (dotted black), and their intersection (whose abscissa is the estimate of the band gap) illustrate the band gap estimation procedure.

DA band gap and a single IA band gap for each of the 49 duplicate samples; and the corresponding band gap energies are shown using false color scale representation in Figure 4a,b, respectively. The mean DA band gap energy (E_g^{DA}) is 2.372 eV, and the mean IA band gap energy (E_g^{IA}) is 2.062 eV with standard deviations of 0.018 and 0.002 eV, respectively. Out of the set of 49 samples, the *TP* spectra corresponding to the smallest and largest estimates of DA and IA gaps are shown in Figure 4c,d. These figures demonstrate that the small variations in the calculated DA and IA band gap energies are a result of variations in the experimental data rather than the performance of the algorithm. In particular, while the different shapes of the TP^{DA} spectra in Figure 4c result in substantially different linear segmentations, the extracted band gap energy is consistent due to the judicious selection of the Tauc and baseline line segments. The 49 independent measurements with automated band gap calculations consistently produced DA and IA band gap energies that are in excellent agreement with previously reported optical characterization of α-Fe_2O_3 (E_g^{DA} approximately 2.2 eV and E_g^{IA} in the range 1.8−2.0 eV).[23]

As a second example of band gap estimation from *TP* spectra obtained from the TR instrument, we apply our algorithm to a sputtered thin film sample of α-$Cu_2V_2O_7$, a recently reported indirect-gap photoelectrocatalyst.[24] Figure 5 demonstrates the estimation of the DA and IA band gap energies as 2.43 and 2.06 eV, respectively. The Tauc analysis of the IA gap in Figure 5 provides an illuminating example of the importance of using a slightly negative value for the minimum baseline slope parameter p_5 in eq 9c. For the TP^{IA} signal, an expert analyst would likely draw a zero-slope line that extends from the minimum value of the TP^{IA}. The linear segment that best simulates such a baseline was correctly selected by the automated algorithm, and this line segment happens to have a slightly negative slope. The resulting estimation of band gap energy is in excellent agreement with that of an expert analyst.

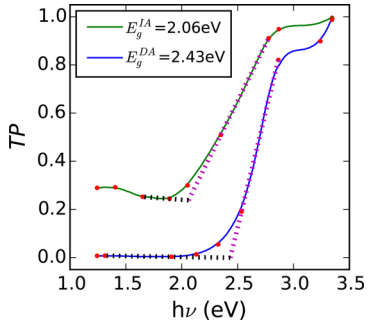


Figure 5. Direct-allowed (DA) and indirect-allowed (IA) *TP* spectra for α-$Cu_2V_2O_7$. End points of the piecewise linear fit (red circles), Tauc line segments (dotted magenta), and baseline segments (dotted black) and the estimates of E_g^{DA} and E_g^{IA} are shown.

To demonstrate band gap estimation on data obtained from the DR instrument, we apply our algorithm to TP^{DA} spectra of representative samples from a sputter-deposited composition library.[25] This library will be discussed further in future work, and for the present purposes, we select representative samples, starting with two samples identified through X-ray diffraction measurements to have the monoclinic $BiVO_4$ structure. As shown in Figure 6, we observe very precise estimation of the DA gap, in excellent agreement with the literature value of 2.49 eV.[26]

Another illustrative example from the (Bi−V−Fe)O_x library is shown in Figure 7. This sample is more Bi-rich and contains two phases, including the same $BiVO_4$ phase as the samples in Figure 5. As noted in the algorithm, the linear segmentation of a given *TP* can be analyzed for the presence of multiple band gaps. For this biphasic sample, the $BiVO_4$ direct band gap is successfully identified, and a second DA band gap near 3.0 eV is also identified. The simultaneous measurement of two band gap energies cannot be as robust as the individual measurement of two phase-pure samples, but for combinatorial band gap

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673−681

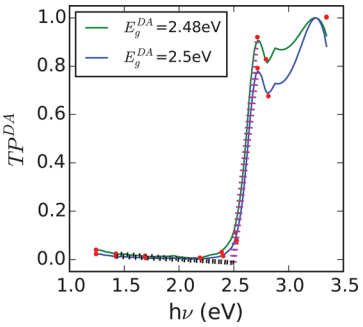


Figure 6. Direct-allowed (DA) *TP* spectra for two monoclinic $BiVO_4$ samples characterized using the diffuse reflectance (DR) technique. For each sample, end points of the piecewise linear fit (red circles), Tauc line segments (dotted magenta), baseline segments (dotted black), and the estimate of E_g^{DA} are shown.

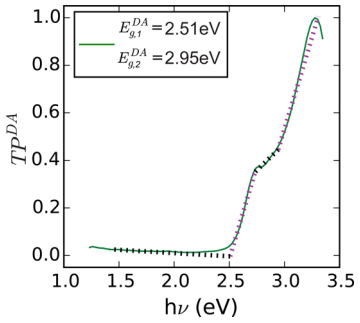


Figure 7. Direct-allowed (DA) *TP* spectra for a biphasic sample in which two DA gaps are identified. Both Tauc line segments (dotted magenta) and baseline segments (dotted black) are shown.

mapping, the ability to identify multiple band gaps is useful for studying composition−property relationships. Since the identification of two band gaps indicates that the given sample is not phase-pure, this result has implications for the underlying phase diagram, which is generally not known for a combinatorial material library. Algorithms for interpreting composition−band gap trends is the subject of ongoing research, and the automated Tauc analysis described in the present work provides a crucial capability for enabling high throughput materials discovery.

It is worth noting that the automated analysis of the $BiVO_4$ samples also identifies an indirect-allowed (IA) gap at 2.44 eV (see Figure S1) that we disregard due to its proximity to the direct-allowed transition, which typically generates much stronger absorption. That is, an IA band gap can only be reliably identified using Tauc analysis when the corresponding Tauc line segment does not strongly overlap with the onset of DA absorption. The materials and data of Figures 4 and 5 exhibit energy differential values ($E_g^{DA} - E_g^{IA}$) of approximately 0.3 and 0.4 eV, respectively. These examples provide an important rule of thumb that an IA band gap can only be detected by absorption spectroscopy and automated Tauc analysis if the IA band gap energy is separated from the lowest DA transition by approximately 0.3 eV or more. For discovery of photoabsorbers, IA transitions are of primary interest when the direct band gap is at substantially higher energy, in which case the significantly lower indirect-allowed band gap can provide insight into composition-dependent phenomena such as band gap tuning.[27] Thus, for the purpose of photoabsorber discovery, Figures 4−7 provide excellent demonstration of the precise identification of DA and IA band gaps using the automated Tauc analysis algorithm.

To demonstrate the algorithm's ability to estimate a range of band gap energies for various materials whose optical properties are not known, we compare the algorithm-generated band gaps with those estimated by three expert scientists. We generated a representative data set of 60 TP^{DA} spectra (see Figure S2) from our UV−vis characterization database. The corresponding 60 composition samples included metal oxides and metal sulfides with a total of 11 cation elements represented and up to four cations present in each sample. The data set consisted of 50 TP^{DA} spectra for which the algorithm identified a single band DA band gap via the $\mathbf{C}_1$ criteria and 10 spectra for which a DA band gap was not identified. For each spectrum, the ground truth for the presence/absence of a DA band gap was established by majority consensus of the three expert scientists. The performance of the automated algorithm in mimicking the expert judgment is shown in Figure 8a, revealing excellent true-positive (98%) and true-negative (82%) rates for spectra with ground-truth presence and absence of a DA band gap, respectively. The algorithm and parameter values (Table 1) were intentionally chosen to minimize the false-negative rate at the expense of the false-positive rate, as desired for material discovery applications. The false-positive rate can be lowered if so desired by tuning parameters to make the Tauc and baseline identification criteria more restrictive. Indeed, the extent to which a given *TP* exceeds or fails inequalities in the criteria sets could be used to ascertain a confidence level in the positive or negative band gap detection, respectively. The goodness of fit of the linear segment model, as quantified by the loss function used in the fitting routine (or similar metrics), could also be used to ascertain uncertainty (or lack of confidence), although

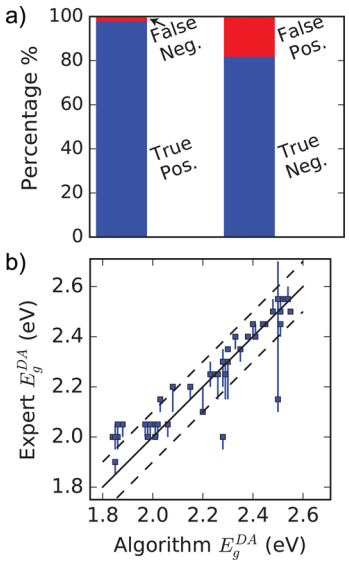


Figure 8. (a) True-positive, true-negative, false-positive, and false-negative percentages for the automated algorithm's ability to identify a band gap given the presence/absence of a band gap as per the majority consensus of three expert scientists as ground truth. (b) Comparison of band gap energies estimated by expert scientists and by the automated algorithm for the 48 true-positive samples. The range of values (for each sample) reported by the three expert scientists is denoted by a vertical line with the median value denoted by a blue marker. A solid black line representing ideal match and dashed black lines representing a mismatch of 0.1 eV are also shown.

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673−681

in the spirit of mimicking expert judgment, we consider the difference in slope $S_\tau - S_B$ to be the most straightforward indicator of confidence in the presence of a band gap. While the comparison to expert results included only two false-positive cases, Figure S3 shows the automated Tauc construction for these spectra corresponded to relatively small values of $S_\tau - S_B$, which supports the use of this metric as an indicator of confidence in the band gap identification.

For the true-positive samples, Figure 8b shows a comparison of the band gap energies estimated by the expert scientists and by the automated algorithm, demonstrating that the discrepancies are typically within 0.1 eV, with an R^2 value of 0.89 between the mean of expert scientist estimated band gaps and algorithm estimated band gap. For several TP^{DA} spectra with band gap energy near 1.8 eV, we observe that modeling of the baseline signal is unusually subjective since the transient in TP^{DA} extends beyond the lower energy limit of the spectra (1.3 eV), resulting in more significant discrepancies in this region of Figure 8b. This provides an important guide for experiment design that the lower and higher energy limit of the spectra should extend at least 0.5 eV beyond the band gap energy range of interest. Figure 8b also contains two outlier data points, which upon subsequent inspection appear to be the result of unusually subjective analysis due to the convolution of multiple overlapping DA transitions in each TP^{DA} spectrum (see plots 14 and 33 in Figure S2).

While these results demonstrate the successful automation of band gap extraction, we also note that the algorithm is based on parameters that are intuitive to an expert scientist, facilitating adaption of the algorithm for different instruments or particular research purposes. We also find that the algorithm enables quality control and exploration of various optical properties through monitoring the following algorithm outputs: slope of the Tauc line segment, slope of the background line segment, criterion set used for identifying the Tauc line segment, number of band gaps identified, and the area under the Tauc line segment. Finally, deployment of this algorithm in high throughput experiments will provide training data for machine learning algorithms to infer and predict composition−structure−property relationships.

SUMMARY

High-throughput mapping of optical properties is of significant importance for rapid discovery of materials for a variety of applications, especially solar energy technologies. In this article, we develop an analysis methodology that allows us to rapidly estimate band gap energy and demonstrate the utility of the algorithm by estimating band gap energies for α-Fe_2O_3, α-$Cu_2V_2O_7$, and monoclinic-$BiVO_4$ phases. While the algorithm has several free parameters, we used expert researchers to train the values of these parameters and demonstrate that the presented values yield an algorithm with excellent reproducibility and ability to estimate band gap energies from both diffuse reflectance and transmission−reflection measurements. The intuitive nature of the free parameters allows further fine-tuning of the accuracy and tailoring to specific experiments when necessary.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscombsci.6b00053.

TP^{DA} plots for 60 samples used for expert−automated algorithm comparison, identification of false positives using descriptors obtained from the automated algorithm, and guidelines used by expert scientists for band gap estimation (PDF)

Data used to generate figures and source code for Python 2.7 implementation of the algorithm (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: gregoire@caltech.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is performed by the Joint Center for Artificial Photosynthesis, a DOE Energy Innovation Hub, supported through the Office of Science of the U.S. Department of Energy under Award Number DE-SC0004993. The authors thank Earl Cornell and Slobodan Mitrovic for assistance with instrument hardware and initial efforts in data processing, Thomas F. Jaramillo for providing insight into the rigors of band gap estimation using Tauc plots, Meyer Pesenson for helpful discussions with data processing, and Lan Zhou for assistance with sample preparation.

REFERENCES

(1) (a) Perkins, J. D.; Paudel, T. R.; Zakutayev, A.; Ndione, P. F.; Parilla, P. A.; Young, D. L.; Lany, S.; Ginley, D. S.; Zunger, A.; Perry, N. H.; Tang, Y.; Grayson, M.; Mason, T. O.; Bettinger, J. S.; Shi, Y.; Toney, M. F. Inverse design approach to hole doping in ternary oxides: Enhancing p-type conductivity in cobalt oxide spinels. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2011**, *84* (20), 205207. (b) Barron, S. C.; Gorham, J. M.; Green, M. L. Thermochromic Phase Transitions in VO2-Based Thin Films for Energy-Saving Applications. *ECS Trans.* **2014**, *61* (2), 387−393. (c) Kirby, S. D.; van Dover, R. B. Improved conductivity of ZnO through codoping with In and Al. *Thin Solid Films* **2009**, *517*, 1958−1960.

(2) Perkins, J. D.; Teplin, C. W.; van Hest, M. F.; Alleman, J. L.; Li, X.; Dabney, M. S.; Keyes, B. M.; Gedvilas, L. M.; Ginley, D. S.; Lin, Y.; Lu, Y. Optical analysis of thin film combinatorial libraries. *Appl. Surf. Sci.* **2004**, *223* (1−3), 124−132.

(3) Hu, S.; Xiang, C.; Haussener, S.; Berger, A. D.; Lewis, N. S. An analysis of the optimal band gaps of light absorbers in integrated tandem photoelectrochemical water-splitting systems. *Energy Environ. Sci.* **2013**, *6* (10), 2984.

(4) Yin, Z.; Tang, X. A review of energy bandgap engineering in III−V semiconductor alloys for mid-infrared laser applications. *Solid-State Electron.* **2007**, *51* (1), 6−15.

(5) Mitrovic, S.; Cornell, E. W.; Marcin, M. R.; Jones, R. J.; Newhouse, P. F.; Suram, S. K.; Jin, J.; Gregoire, J. M. High-throughput on-the-fly scanning ultraviolet-visible dual-sphere spectrometer. *Rev. Sci. Instrum.* **2015**, *86* (1), 013904.

(6) (a) Wood, D.; Tauc, J. Weak Absorption Tails in Amorphous Semiconductors. *Phys. Rev. B* **1972**, *5*, 3144−3151. (b) Davis, E. A.; Mott, N. F. Conduction in non-crystalline systems V. Conductivity, optical absorption and photoconductivity in amorphous semiconductors. *Philos. Mag.* **1970**, *22*, 0903−0922.

(7) (a) Tauc, J.; Grigorovici, R.; Vancu, A. Optical Properties and Electronic Structure of Amorphous Germanium. *Phys. Status Solidi B* **1966**, *15*, 627−637. (b) Tauc, J.; Menth, A.; Wood, D. L. Optical and Magnetic Investigations of the Localized States in Semiconducting Glasses. *Phys. Rev. Lett.* **1970**, *25* (11), 749−752.

(8) Escobedo Morales, A.; Sanchez Mora, E.; Pal, U. Use of diffuse reflectance spectroscopy for optical characterization of un-supported nanostructures. *Rev. Mex. Fis.* **2007**, *53* (5), 18−22.

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673−681

(9) Dolgonos, A.; Mason, T. O.; Poeppelmeier, K. R. Direct optical band gap measurement in polycrystalline semiconductors: A critical look at the Tauc method. *J. Solid State Chem.* **2016**, *240*, 43−48.
(10) (a) Moss, T. S. Theory of the Spectral Distribution of Recombination Radiation from InSb. *Proc. Phys. Soc., London, Sect. B* **1957**, *70* (2), 247−250. (b) Burstein, E. Anomalous Optical Absorption Limit in InSb. *Phys. Rev.* **1954**, *93* (3), 632−633.
(11) Berggren, K. F.; Sernelius, B. E. Band-gap narrowing in heavily doped many-valley semiconductors. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1981**, *24* (4), 1971−1986.
(12) (a) Asahi, R.; Morikawa, T.; Ohwaki, T.; Aoki, K.; Taga, Y. Visible-light photocatalysis in nitrogen-doped titanium oxides. *Science* **2001**, *293* (5528), 269−271. (b) Loiudice, A.; Ma, J.; Drisdell, W. S.; Mattox, T. M.; Cooper, J. K.; Thao, T.; Giannini, C.; Yano, J.; Wang, L. W.; Sharp, I. D.; Buonsanti, R. Bandgap Tunability in Sb-Alloyed BiVO4 Quaternary Oxides as Visible Light Absorbers for Solar Fuel Applications. *Adv. Mater.* **2015**, *27* (42), 6733−40.
(13) Anderson, A. Y.; Bouhadana, Y.; Barad, H.-N.; Kupfer, B.; Rosh-Hodesh, E.; Aviv, H.; Tischler, Y. R.; Rühle, S.; Zaban, A. Quantum efficiency and bandgap analysis for combinatorial photovoltaics: sorting activity of Cu-O compounds in all-oxide device libraries. *ACS Comb. Sci.* **2014**, *16*, 53−65.
(14) Ghobadi, N. Band gap determination using absorption spectrum fitting procedure. *Int. Nano Lett.* **2013**, *3*, 2.
(15) (a) Drobny, V. F.; Pulfrey, L. Properties of reactively-sputtered copper oxide thin films. *Thin Solid Films* **1979**, *61* (1), 89−98. (b) Rakhshani, A. E.; Varghese, J. Optical absorption coefficient and thickness measurement of electrodeposited films of Cu2O. *Phys. Status Solidi A* **1987**, *101* (2), 479−486. (c) Roberts, S. Optical Properties of Copper. *Phys. Rev.* **1960**, *118* (6), 1509−1518.
(16) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36* (8), 1627−1639.
(17) Swinehart, D. F. The Beer-Lambert Law. *J. Chem. Educ.* **1962**, *39* (7), 333.
(18) Peng, H.; Ndione, P. F.; Ginley, D. S.; Zakutayev, A.; Lany, S. Design of Semiconducting TetrahedralMn1−xZnxOAlloys and Their Application to Solar Water Splitting. *Phys. Rev. X* **2015**, *5* (2), 021016.
(19) Kubelka, P. New Contributions to the Optics of Intensely Light-Scattering Materials. Part I. *J. Opt. Soc. Am.* **1948**, *38* (5), 448−457.
(20) (a) Murphy, A. B. Band-gap determination from diffuse reflectance measurements of semiconductor films, and application to photoelectrochemical water-splitting. *Sol. Energy Mater. Sol. Cells* **2007**, *91*, 1326−1337. (b) Murphy, A. B. Optical properties of an optically rough coating from inversion of diffuse reflectance measurements. *Appl. Opt.* **2007**, *46*, 3133−43.
(21) Fujiwara, H. *Spectroscopic Ellipsometry: Principles and Applications*; John Wiley & Sons: Hoboken, NJ, 2007.
(22) Hishikawa, Y.; Nakamura, N.; Tsuda, S.; Nakano, S.; Kishi, Y.; Kuwano, Y. Interference-Free Determination of the Optical Absorption Coefficient and the Optical Gap of Amorphous Silicon Thin Films. *Jpn. J. Appl. Phys.* **1991**, *30* (Part 1, No. 5), 1008−1014.
(23) Al-Kuhaili, M. F.; Saleem, M.; Durrani, S. M. A. Optical properties of iron oxide (α-Fe2O3) thin films deposited by the reactive evaporation of iron. *J. Alloys Compd.* **2012**, *521*, 178−182.
(24) Zhou, L.; Yan, Q.; Shinde, A.; Guevarra, D.; Newhouse, P. F.; Becerra-Stasiewicz, N.; Chatman, S. M.; Haber, J. A.; Neaton, J. B.; Gregoire, J. M. High Throughput Discovery of Solar Fuels Photoanodes in the CuO−V2O5 System. *Adv. Energy Mater.* **2015**, *5*, 1500968.
(25) Suram, S. K.; Zhou, L.; Becerra-Stasiewicz, N.; Kan, K.; Jones, R. J. R.; Kendrick, B. M.; Gregoire, J. M. Combinatorial thin film composition mapping using three dimensional deposition profiles. *Rev. Sci. Instrum.* **2015**, *86* (3), 033904−033904.
(26) Payne, D. J.; Robinson, M. D. M.; Egdell, R. G.; Walsh, A.; McNulty, J.; Smith, K. E.; Piper, L. F. J. The nature of electron lone pairs in BiVO4. *Appl. Phys. Lett.* **2011**, *98* (21), 212110.
(27) Copple, A.; Ralston, N.; Peng, X. Engineering direct-indirect band gap transition in wurtzite GaAs nanowires through size and uniaxial strain. *Appl. Phys. Lett.* **2012**, *100* (19), 193108.

DOI: 10.1021/acscombsci.6b00053
ACS Comb. Sci. 2016, 18, 673−681